Volume Title: Frontiers in Health Policy Research, Volume 4

Volume Author/Editor: Alan M. Garber, editor

Volume Publisher: MIT PRess

Volume ISBN: 0-262-57150-1

Volume URL: http://www.nber.org/books/garb01-1

Publication Date: January 2001

Chapter Title: Creating Price Indexes for Measuring Productivity in Mental Health Care

Chapter Author: Susan H. Busch, Ernst R. Berndt, Richard G. Frank

Chapter URL: http://www.nber.org/chapters/c9854

Chapter pages in book: (p. 115 - 147)

5

Creating Price Indexes for Measuring Productivity in Mental Health Care

Susan H. Busch, *Division of Health Policy and Administration, Yale Medical School*
Ernst R. Berndt, *Alfred P. Sloan School of Management, MIT and NBER*
Richard G. Frank, *Department of Health Care Policy, Harvard Medical School and NBER*

Executive Summary

Economists have long suggested that to be reliable, a preferred medical care price index should employ time-varying weights to measure outcomes-adjusted changes in the price of treating an episode of illness. In this article, we report on several years of research developing alternative indexes for the treatment of the acute phase of major depression, for the period 1991–1996. The introduction of new treatment technologies in the past two decades suggests well-known measurement issues may be prominent in constructing such a price index.

We report on the results of four successively refined methods to measure price changes. In contrast with Bureau of Labor Statistics (BLS) price indexes, we find prices decline and productivity increases. In Method 1, we employ a publicly available retrospective claims database to create BLS-like price indexes for two discrete inputs used in the treatment of depression: antidepressant drugs and physician's services. In Method 2, we move to an episode-based price index. As a first attempt to adjust for outcome variations, we limit the analysis to episodes receiving one of seven guideline-level treatments, each of which have been shown to have equivalent efficacy in clinical trials. In Method 3, we expand the number of episodes included and allow for finer adjustments in expected outcomes. We assign expected outcomes to episodes based on an expert panel's rating of the effectiveness of the treatment received. In Method 4, our preferred price index, we use hedonic regression to adjust also for changes in patient characteristics that add to treatment complexity.

We find that our preferred method suggests a *decline* in the AAGR of treatment for depression of 2.7 percent, which is in contrast to a price *increase* of 2.6 percent per year when using BLS-like methods. This research suggests constructing episode-based, outcomes-adjusted price indexes is a complex and cumbersome task but an important one for public policy decisions. Although it may not be sensible or practical for the BLS to produce such an index on a monthly basis, it is important that policy analysts use episode-based out-

comes-adjusted price indexes when evaluating productivity based on National Health Accounts.

I. Introduction

Measuring price changes and productivity is important in health care just as it is in the rest of the economy. More complicated, however, is the concept of output in the health care sector. For decades, economic statisticians and health economists have wrestled with various ways of characterizing output for the purposes of assessing prices and productivity in health care. Productivity measurement takes on special significance in the health care sector because various well-known market failures preclude one from confidently making normative inferences based on observed market outcomes. For this reason, productivity measurement and the development of price indexes provide an important source of information on the "value" of health care services.

Nowhere in the health care sector are the problems of assessing the value of treatments and productivity of spending more challenging than in the case of treatments for mental disorders. Questions about the effectiveness of treatments and welfare losses from moral hazard in insurance have long created concern that the value of spending on mental health may be low relative to other health services. Because mental disorders are often chronic, recurring conditions, mortality is not typically an appropriate outcome measure. Thus, defining outcomes often relies on more subjective and difficult to measure constructs. Creating price indexes that account for the changing quality and effectiveness of mental health treatments therefore poses important measurement issues.

For some time, health economists have posited that an appropriate price index is one based on episodes of treatment of selected illnesses and conditions, incorporating technological and institutional innovations that change the mix of inputs used to treat the condition, and including any effects on changed medical outcomes. Scitovsky (1967) illustrated one approach to implementing an episodes-based analysis of prices by examining the changing cost of treating episodes of six specific illnesses. The research line described below builds on the tradition begun by Scitovsky and applies it to the most prevalent and costly of the mental disorders, major depression. In this article, we report on several years of research focused on developing price indexes for the treatment of depression.

Several features of treatment for depression raise well-known measurement issues that may be prominent in constructing such a price index. In the past two decades, new technologies have been introduced indicating the potential for changes in outcomes of treatment. Treatment patterns have shifted within treatment classes (e.g., from older to more recently developed drugs) and between treatment classes (e.g., from psychotherapy to drug treatments). Fundamental organizational changes may also have affected prices and treatment choice.

We begin by describing the methods used by the BLS to construct official health-related Producer Price Indexes (PPIs). We then describe four methods for constructing a supply price index for treatment of depression. In Method 1, we mimic the procedures used by the BLS but employ prices and quantities from a publicly available private insurance claims database. We also deviate from BLS procedures by using the claims data in conjunction with index number formulas not currently used by the BLS. In the three subsequent methods, we take an episode of illness approach, pricing treatment for a specific illness episode rather than discrete inputs. Method 2 aims to hold outcome constant by limiting the price index to episodes of treatment that meet a common outcome standard, as determined by the outcomes of patients assigned these treatments in clinical trials. Method 3 incorporates variation in expected clinical outcomes by assigning expected outcomes to treatments based on ratings of treatments by a panel of experts. Method 4 uses the same methods as 3, but adds a regression analysis, allowing us to adjust prices for varying patient characteristics that may complicate treatment.

In our first analysis, we apply BLS methods to claims data. The results based on the BLS methods allow us to determine whether our results obtained using alternative methods, such as the "episodes" approach, are due to differences in data or differences in the methods. Unlike the data collected by the BLS, the data we employ is from a nonrepresentative sample of privately insured individuals with extensive health insurance coverage. In addition, these data include transactions rather than list prices. This distinction may be important if organizational changes in health insurance during the time period considered have resulted in the negotiation of larger discounts, thus increasing the difference between transaction and list prices over time.

For Methods 2, 3, and 4, we move from pricing discrete inputs to pricing an episode of treatment. We use health plan enrollment information, diagnoses and dates contained in the retrospective claims data to combine claims such that they reflect episodes of outpatient

treatment for major depression. These treatment episodes consist of clinician visits of various types, pharmaceuticals, or combinations of these inputs. Information on outcomes is not recorded in administrative data, which necessitates a second source for outcome information. Based on a review of the clinical literature and published treatment guidelines, we identify in Method 2 all episodes treated with proven efficacious treatments. To increase the validity of this method, we make considerable efforts to match closely the treatments reflected in the data to treatments evaluated in clinical trials.

By limiting the price index to episodes treated with guideline-level treatment, we omit about 50 percent of the observed episodes from the price index. We also implicitly assume that all guideline-level treatments produce similar clinical benefits, an unlikely prospect. Descriptive results indicate an increase over time in the share of episodes treated with guideline-level treatment between 1991 and 1996 (from 35 to 55 percent). A real source of productivity increase is thereby overlooked by applying Method 2. In Method 3, we allow outcomes to vary by treatment and patient type. Allowing this variation increases the number of episodes included in the price index. We expand the use of information on outcomes by convening a panel of experts, who were asked to estimate the expected outcomes of treatments reported in the data. These ratings estimates were then combined with price and quantity data to construct price indexes that take into account expected outcomes.

In Method 4, we maintain the same outcome measurement approach and episode construction as in Method 3, but we recognize that the patient population may have been changing over time. Method 4 uses hedonic regressions to adjust for demographic shifts and the changing complexity of treated cases. This adjustment is accomplished by measuring the presence of co-occurring medical and psychiatric conditions that complicate care and tend to diminish responses to treatment.

The procedures and data underlying Methods 1 through 4 are summarized in table 5.1. In contrast with official price indexes, the price indexes we construct suggest real price declines, although the magnitude varies across methods. We focus on what has been learned about constructing a preferred medical care price index both generally and specifically for the treatment of depression.

In the next section, we discuss major depression and its treatment. Section III describes the data used in the analyses. Section IV discusses methods used by BLS to construct official price indexes. Section V de-

Table 5.1
Summary of alternative price index calculation methods

	BLS PPI	Method 1	Method 2a	Method 2b	Method 3	Method 4
Data	National sample	MarketScan	MarketScan	MarketScan	MarketScan	MarketScan
Missing psychotherapy codes identified	NA	Yes	No	Yes	Yes	Yes
Limit to first four months, or acute phase, of treatment	NA	No	No	Yes	Yes	Yes
Years	1991–1995	1991–1995	1991–1995	1991–1995	1991–1996	1991–1996
Unit of analysis	Discrete inputs	Discrete inputs	Episode of treatment	Episode of treatment	Episode of treatment	Episode of treatment
Limit analysis to episodes with equivalent efficacy, based on results in clinical trials	No	No	Yes, also exclude mixed drug treatments	Yes	No	No
Assign outcomes based on expert panel ratings	No	No	No	No	Yes	Yes
Percent of all episodes included in the price index	NA	NA	25	50	67	67
Adjust patient base for complications	No	No	No	No	No	Yes

scribes Method 1. Section VI describes the construction of episodes and the procedures used in Method 2 to assign outcomes based on clinical trial results. The quantities and prices of guideline-level treatment bundles are also described. Section VII describes Method 3 and the elicitation methods used to gather data on outcomes from an expert panel, and incorporates these outcomes into a price index. In Section VIII, we incorporate changes in the characteristics of the patients used in the calculation of the Method 4 price indexes. In Section IX, we consider the episodes-based price indexes in the context of official price indexes, decompose the discrepancies, and interpret them. Section X presents a conclusion.

II. Depression

Epidemiological research indicates that in the early 1990s, 10.3 percent of the U.S. population met the criteria for major depression at some time during a twelve-month period.[1] Depression is often accompanied by health problems. They include anxiety, eating disorders, substance abuse, and other medical conditions.[2] Although the reasons are still not well understood, women—particularly women under age 25—are much more likely to suffer from depression than are men; the relative lifetime female/male prevalence rate is about 1.7. Rates of recurrence for major depression appear to be no different for women and men.[3] Several studies have shown that depression has similar or greater functional impairments than those attributed to other episodic and chronic medical illnesses.

Several new treatments for depression have been introduced in the past two decades. In the area of psychotherapy, various new techniques have expanded treatment options, including interpersonal psychotherapy (IPT), family therapy, and cognitive behavioral therapy (CBT). In addition, extraordinary advances have been achieved in the area of antidepressant medication. Antidepressant medication has three general classes: tricyclic antidepressants (TCAs); selective serotonin reuptake inhibitors (SSRIs), which include brand-name drugs such as Prozac, Zoloft, and Paxil; and monoamine inhibitors (MAOIs), which, due to side effects and dangerous interactions, are generally used only for cases that are resistant to other forms of treatment. The newer SSRIs offer some distinct advantages over older TCAs. SSRIs are associated with lower risk of overdose and fewer and lower levels of

several side effects. The advantages of SSRIs come at a significantly higher pecuniary cost than most TCAs.

A review of approximately thirty major clinical trials and meta-analyses from the clinical literature on comparative efficacy of acute phase treatments points to several key conclusions.[4] First, psychotherapies of all kinds have been shown to result in superior outcomes compared to no treatment. When compared amongst themselves, the different forms of psychotherapy appear to have no significant differences in outcomes.[5] Second, for less severe forms of depression, psychotherapies alone, TCAs with medical management, and SSRIs with medical management appear to produce comparable outcomes. Each of these therapies produce significantly better outcomes than placebo treatments. Combination treatments with these therapies as components also generate equivalent levels of efficacy for less severe forms of depression.

Third, for more severe forms of depression, the bulk of the evidence suggests that TCAs alone, SSRIs alone, and combinations of drugs and psychotherapy have comparable levels of efficacy and that each results in superior outcomes compared to psychotherapy alone. Recently some evidence has emerged showing some extra improvement from the combination treatments relative to medication alone,[6] particularly for more chronic forms of depression.[7]

III. Data

In all cases, approaches to price index construction are compared using quantities and prices of outpatient treatment for depression based on retrospective claims data from the publicly available MarketScan™ database. These data consist of 1991–1996 medical claims from four large self-insured employers that offered more than twenty-five health plans to 428,000 employees and their dependents. The data consist of enrollment, inpatient, outpatient, and pharmaceutical claims. This data may not be representative of the prices paid or treatment provided by other purchasers such as individuals, government, or small employers, thereby limiting the generalizability of some of our findings. For example, the health benefits offered to enrollees in this database are quite generous relative to the general market for private health insurance in the United States. Individuals insured by private payers may be healthier than the general population, suggesting less complicated treatment

plans. Large employers may have the market power to negotiate substantial price discounts, reducing prices compared with the average purchaser.

IV. BLS Procedures

Although the Consumer Price Index (CPI) is generally the most widely reported measure of changes in general prices, in this article we focus on comparisons with Producer Price Indexes (PPIs). Medical-related CPIs limit their scope to consumers' out-of-pocket payments and that portion of third-party insurance paid for out-of-pocket by employees. Employers' contributions to employee health insurance are excluded, as are all medical payments to providers by governments. Thus, although national health spending in the mid-1990s constituted about 14 percent of GDP, the total weight given medical care items in the CPI was only about 5 percent.[8] These limits on the definition of "consumer price" cloud the interpretation of the CPI for health care. Unlike the CPI, medical PPIs measure average changes in selling prices received by domestic producers for their output and include revenues from Medicare, Medicaid, and other private sources (such as third-party payment and direct patient cash payments), making the PPI a more reasonable comparison for examining changes in medical care prices.

For quite some time, the BLS has published PPIs for certain medical-related manufacturing industries, such as pharmaceuticals and diagnostic equipment. Only relatively recently, however, has the BLS begun publishing PPIs for medical service industries such as hospital and physician services. The BLS initiated a PPI for an aggregate of health care services in 1994, for offices and clinics of doctors of medicine in 1993, and for hospitals in 1992. These PPIs encompass various more detailed industries. For example, offices and clinics of doctors of medicine includes physician services from psychiatrists, general/family practitioners, and internists. Hospitals includes hospital services from general medical and surgical hospitals, psychiatric hospitals, and specialty non-psychiatric hospitals. Select PPIs relevant to mental health care are listed in table 5.2. These medical care PPIs indicate price increases ranging from 2.2 percent to 6.1 percent per year in the early to mid-1990s. Of the medical PPIs published by the BLS, the two that are most relevant to outpatient care for the acute phase of major depression are psychotherapeutic drugs and physician services (offices of doctors). While we focus here on these two PPIs, many of the

Table 5.2
Mental health related BLS Producer Price Indexes

Producer category	1991	1992	1993	1994	1995	1996	AAGR
Office of doctors	—	—	100	102.8	105.9	107.8	+0.025
Psychiatrists, small group/solo	—	—	100	102.9	104.7	106.8	+0.022
All outpatient services	—	—	100	102.4	106.9	114.2	+0.045
Psychiatric hospital outpatient services	—	—	100	103.3	113.0	119.4	+0.061
Psychotherapeutics	100[a]	107.6	113.4	116.2	120.4	120.5	+0.038

[a]Index normalized to 1991 base year.
Source: U.S. Bureau of Labor Statistics, Producer Price Indexes.

conceptual and empirical issues discussed are common to other Producer Price Indexes.

Pricing the output of prescription pharmaceuticals presents some particularly interesting issues involving treatment of generic drugs and of quality improvements in new products that are beyond the scope of this article.[9] For our purposes, it is important to note that to construct pharmaceutical PPIs, the BLS chooses a sample of drugs that is then repriced each month. With respect to physician services, the BLS constructs a national sample of physician practice units from which a random bill is chosen. From this bill, the BLS calculates the net price paid to the practice for the entire set of services and procedures provided during a single office visit, distinguished by payer type. The physician's output from this visit is represented by the content of the patient's bill, including specific procedure (CPT-4) codes associated with that visit, and is related to a particular ICD-9 diagnosis. Given this sample bill, the BLS contacts the physician practice unit each month and asks it to reprice what the current net transactions prices would be for that particular fixed bundle/payer of services. Hence, the unit of analysis or industry output is the group of procedures delivered during the randomly selected single office visit.

It is worth noting that, given these definitions of industry output and the BLS' use of fixed itemized components for obtaining price quotes, major input substitution of treatment for a condition (such as the changing mix of outpatient and inpatient care or, in the case of outpatient care for major depression, psychotherapy and psychotherapeutic

drugs) is not captured. Such a zero-substitutability definition of medical outputs leads to the existence of a substitution bias.

Although the BLS seeks transactions rather than list prices for its price quotes, the agency is aware that compliance by firms is easier with list than with actual transaction prices. The BLS currently draws a sample of items for each industry on average about every seven years and then reprices this fixed set of items monthly until an entirely new sample is drawn. In recognition of the fact that seven-year time lags could yield a sample of products and services unrepresentative of current market transactions, the BLS announced in 1995 that samples would be supplemented at one- or two-year intervals for certain technologically dynamic industries, including pharmaceuticals. PPIs are then constructed using a modified fixed weight Laspeyres price index formula over the individual price quotes.

Finally, these BLS price index procedures do not take into account any variations across therapies and over time in expected outcomes from treatment. Further discussion of the BLS' procedures for measuring prices in the health care industries is given in Berndt, Cutler, Frank et al. (2000b).

V. Method 1: An Index That Mimics the BLS

To distinguish differences arising from the methods used to construct the price index from differences due to data, we construct a price index that mimics the BLS' methods but uses as inputs price and quantity data from the MarketScan data. As discussed above, MarketScan data differs from data collected by BLS because patients are from a nonrepresentative subset of the population, transaction rather than list prices are recorded, and prices for all outpatient services and pharmaceuticals are utilized instead of those from a random sample. We initially focus on PPIs for two discrete inputs commonly used in the treatment of major depression: psychotherapeutic drugs and physician services.

Psychotherapeutic pharmaceuticals used in the treatment of depression are identified from national drug codes (NDC) reported on pharmaceutical claims. For a claim to be included, the drug had to be FDA-approved for the treatment of depression and had to be linked with a depression diagnosis (ICD-9 codes 296.2 and 296.3). Fourteen distinct molecules were identified.

Physician services for the treatment of depression were identified by procedure codes (CPT-4) attached to a major depression diagnosis in

the outpatient claims of the MarketScan data.[10] In all years, the most common procedure was a fifty-minute psychotherapy visit. Other procedures included lab tests, psychiatric diagnostic testing, general office visits, and medical management visits.

Following BLS methods, the fixed quantity weight Laspeyres price index formula was used to aggregate prices and quantities. Various other price indexes can be computed, each having implicit assumptions on the extent of ex-ante substitutability among the alternative treatment bundles. Because alternative index number formulas and underlying assumptions are well known in the literature,[11] and since in our research we found choice of formula not to affect price change greatly, we present only two price indexes here: a fixed weight Laspeyres and a Fisher-ideal index. The Fisher-ideal index is a geometric mean of the fixed-weight Laspeyres and fixed-weight Paasche indexes, where the latter is analogous to the Laspeyres except that it employs current period quantities rather than base period quantities as fixed weights.

For both the Laspeyres and Fisher-ideal price indexes, we construct an aggregate price index for antidepressant drugs and for physicians' services using the Medstat MarketScan™ data. The top row of tables 5.3 and 5.4 presents the BLS-type antidepressant price indexes, while that for physicians' services is given in the second row of tables 5.3 and 5.4. For antidepressant drugs, both the fixed weight Laspeyres and Fisher-ideal index rise at an average annual growth rate (AAGR) of 4.1 percent, slightly greater than the 3.8 percent in the BLS' PPI for psychotherapeutic drugs. Hence, for antidepressant drugs, there is relatively little difference in the BLS' index and that based on the Medstat data.

With respect to physicians' services, as seen in the second row of table 5.3, the Laspeyres index based on Medstat data has an AAGR of 2.5 percent, a rate identical to that for the BLS' PPI for office of doctors (table 5.2). When one employs instead the Fisher-ideal price index, the physician service price index grows less rapidly, at 1.8 percent per year. This lower growth rate could reflect the increasing role over time of general/family practitioners and internists in treating depression and the reduced role of psychiatrists, a treatment shift not captured by the fixed weight Laspeyres index.

In summary, the AAGRs of price indexes based on the fixed weight Laspeyres index, using Medstat data, are very similar to the BLS' PPIs for antidepressant drugs and for physicians' services, based on their own sample data. Hence, the underlying BLS and Medstat data are apparently very similar.

Table 5.3
Aggregate Producer Price Indexes for the treatment of acute phase major depression: fixed quantity weight Laspeyres formula

	1991	1992	1993	1994	1995	1996	AAGR
Method 1: Mimic BLS procedures							
BLS-type PPI for antidepressant drugs	100	107.7	108.9	109.4	116.0	122.0	+0.041
BLS-type PPI for physicians' services	100	101.6	101.2	104.5	104.2	113.3	+0.025
Method 2: Limiting to guideline-level treatments							
2a: Prior to revisions to episode construction	100	98.4	86.7	79.2	68.4	—	−0.091
2b: Post revisions to episode construction	100	100.3	97.5	93.1	97.6	—	−0.006
Method 3: Expert panel ratings							
Expand categories with no outcome adjustment	100	97.5	93.7	92.2	95.4	95.0	−0.010
Price per full remission	100	92.0	99.5	101.3	108.3	103.4	+0.007
Method 4: Adjusting for patient characteristics							
Hedonic regression	100	92.3	95.2	93.5	88.0	87.2	−0.027

Table 5.4
Aggregate Producer Price Indexes for the treatment of acute phase major depression: Fisher-Ideal index

	1991	1992	1993	1994	1995	1996	AAGR
Method 1: Mimic BLS procedures							
BLS-type PPI for antidepressant drugs	100	106.2	107.4	108.6	115.8	122.2	+0.041
BLS-type PPI for physicians' services	100	101.4	101.7	103.9	104.3	109.3	+0.018
Method 2: Limiting to guideline-level treatments							
2a: Prior to revisions to episode construction	100	98.3	87.8	80.6	70.1	—	−0.085
2b: Post revisions to episode construction	100	100.7	97.2	94.5	97.3	—	−0.007
Method 3: Expert panel ratings							
Expand categories with no outcome adjustment	100	97.9	95.7	94.6	96.4	97.2	−0.005
Price per full remission	100	92.0	99.8	101.6	107.9	103.7	+0.007

VI. Method 2: Treatment Episodes

Next, we construct price indexes that represent a significant conceptual shift from traditional medical care price indexes. Price indexes that measure the price of discrete inputs, such as those in Method 1, may fail to capture changes in treatment choice that affect productivity. In the approach to price measurement described next, which we refer to as Methods 2a and 2b, we move from pricing inputs in isolation to pricing an episode of treatment.[12] We do this in recognition that a course of treatment more closely reflects the good consumers expect to purchase through medical care—alleviation of disease. We aggregate outpatient and pharmaceutical claims such that combinations of inputs reflect the treatment received by an individual patient. Methods 2a and 2b differ in the algorithm used to construct episodes. Using the episode as the unit of analysis, we aggregate prices and quantities to form price indexes for "the acute phase of treatment for major depression."

Constructing Episodes of Treatment

To implement a price index for treatment episodes, individual claims are combined using patient identifiers, diagnostic information, and dates to represent episodes of treatment. For depression, a chronic disease, defining an acute episode required extensive knowledge of the disease, its course, and the administering of treatments in practice. At several key points, we benefited from consultations with clinicians about these issues.[13] Their input was crucial to developing this approach. We also found the idiosyncrasies of claims data caused additional complexities. To illustrate these difficulties, we describe results from our initial attempt at implementing the episode concept (Method 2a), what we learned, how we improved on these methods (Method 2b), and the changes in the price index resulting from these seemingly minor improvements. While not changing the qualitative findings, the magnitude of our results are affected. In each case, we find declines in real prices, but Method 2b suggested less rapid declines than we originally reported (Method 2a).

We identify all ambulatory claims associated with either single or recurrent episodes of depression, as defined by the ICD-9 primary diagnostic codes 296.2x or 296.3x.[14] We choose these specific diagnosis codes for three reasons. First, clinical trials typically use these definitions as their entry criteria. Because we assign outcomes based on clinical trial results, we want the patients in our data to have symptoms

similar to those patients chosen for clinical trials. Second, chart reviews indicate that the specificity of these two diagnoses is high, i.e., the proportion of true positives is high, while the proportion of false positives is very low. Third, clinicians could employ a more ambiguous diagnosis such as "depression not elsewhere classified" or "neurotic depression." The fact that clinicians designated the diagnosis as either 296.2 or 296.3, rather than codes corresponding to more ambiguous diagnoses, indicates a conscious act of volition.

Since we do not directly observe symptoms in retrospective claims data, our claims-based definition of an episode of treatment does not correspond directly to an episode of the illness.[15] When claims data indicate that psychotherapeutic drugs were prescribed, we consider the number of days of treatment provided by the prescription as the time period over which an individual received care. We follow American Psychiatric Association (1993) guidelines in defining an episode of depression as new if a diagnosis is preceded by a period of at least eight weeks of not meeting clinical criteria for depression.[16] Thus, we use an eight-week period without treatment to define new treatment episodes. We eliminate episodes for which we do not observe the entire episode due to censoring; all episodes for which we do not observe eight weeks both pre- and post-episode are excluded.

We next exclude all depressive episodes involving inpatient (hospital) claims. Patients with only outpatient care constitute the vast majority of treatment episodes (75 to 80 percent). We exclude patients with psychiatric hospitalizations for several reasons. First, inpatient claims data typically do not contain information on the drugs prescribed for treatment; thus, characterizations of inpatient care are inherently incomplete.[17] Second, because of other incomplete information regarding illness severity and comorbid conditions, it is difficult to use administrative claims data to characterize fully an illness diagnosis and therefore to make judgments about the appropriate use of hospital services for treating major depression. Third, few clinical trials specifically address inpatient treatment for major depression, making it difficult for us to assign outcomes to treatments. Finally, because there was considerable evidence of overuse of hospital services in the aggregate during the late 1980s and early 1990s, the inclusion of hospital services in our 1991 base year could make interpretation of price changes troublesome.[18]

Our strategy of limiting severe cases by excluding individuals hospitalized for treatment of depression is likely to be only partly successful. During the 1990s, there has been a substantial reduction in inpatient

psychiatric admissions and in the length of stay for hospitalized cases. An implication of this trend is that the population of people treated only on an outpatient basis may, on average, be suffering from more severe forms of depression over time.

Using information on procedures (e.g., type of visit, whether drug prescribed) as given by the CPT-4 codes, we describe the composition of treatment that occurred within a treatment episode. Drug treatment is based on the national drug codes (NDC) reported on the claim. The NDC classification revealed use of seven TCAs, three SSRIs, two other serotonin-related drugs,[19] three MAO inhibitors, four anxiolytics, and four heterocyclics for treatment of depression. Direct medical spending for such treatment episode was calculated using actual transaction data. All payments made by the insurer to the provider and any cost-sharing assigned to the patient (e.g., patient out-of-pocket copayment for prescription drugs) were summed to a nominal dollar total for each treatment episode.

Assigning Expected Outcome

The results from our review of the clinical trial literature enabled us to develop a set of treatment "bundles" that group therapies into what we envision as therapeutically similar groups for treatment of a specific form of major depression. This identification of treatment bundles that result in similar expected health outcomes is a crucial step in the construction of medical treatment price indexes that employs results from clinical trials to account for expected outcomes. The implicit assumption in this methodology is that obtaining therapeutically similar outcomes from alternative treatments provides a useful approximation to achieving similar expected utility levels. We recognize this is only an approximation. This is particularly true for depression, where the constellation of side effects across treatment can vary significantly and idiosyncratically, and can lead to differential patient compliance and patient preferences between the SSRIs and TCAs.[20]

It is important to note here that our use of guideline standards of care in Method 2 imposes a rather unrealistic shape to the production function for treatment of depression. It takes on a step function form. For example, if one were to receive six psychotherapy visits for treatment of depression, our analysis would treat the case as "effective," whereas four or five visits would be viewed as "ineffective." This is unlikely to provide an accurate representation of clinical reality. Nevertheless, if we limit ourselves to results from the clinical literature, we have little

systematic analysis upon which to make alternative assumptions. Thus, we use the step function production model as a point of departure. In Methods 3 and 4 we relax this assumption.

In Method 2a, we identify five major classes of treatments that have been proven effective in the treatment of depression:

1. Psychotherapy alone, 6–15 visits;
2. Short-term TCA treatment (31–180 days) alone or with medical management;
3. Short-term SSRI treatment (31–180 days) alone or with medical management;
4. Short-term TCA treatment (31–180 days) with some psychotherapy;
5. Short-term SSRI treatment (31–180 days) with some psychotherapy.

Method 2a also employs restrictive inclusion criteria, only considering episodes containing these precise number of visits; for example, episodes with more than fifteen psychotherapy visits are excluded.

Initial empirical analyses suggested several possible improvements to our method of episode construction and censoring. Method 2b incorporates such changes to the data underlying construction of the price indexes. Procedure codes were missing for many outpatient claims. We developed an algorithm to identify missing psychotherapy procedure codes and reconstructed episodes of treatment including these codes. In recognition that in actual practice it is common to switch or augment drug therapies, two treatment categories were added to the five in Method 2a: short-term combined TCA/SSRI treatment, 31–180 days, both alone and in combination with psychotherapy. We also include episodes that involved longer treatment or those extending beyond fifteen visits or six months but count only visits and drugs occurring in the first six months of care. Finally, we improve our definition of censoring, limiting our analysis to the first six months of treatment. Because we believe these procedures represent improvements over Method 2a, we focus our discussion on the results obtained using Method 2b, although we illustrate sensitivity by comparing results from Methods 2a and 2b.

Results

The quantity shares of the seven treatment bundles derived under Method 2b, by year, are presented in table 5.5. Several notable trends appear. First, there has been a very substantial decline in

Table 5.5
Guideline-compatible treatment bundle quantity proportions post revisions to episode construction, by year

Treatment	1991	1992	1993	1994	1995	Overall
Psychotherapy alone	41%	33%	24%	22%	28%	28%
TCA alone	6	5	3	3	2	4
SSRI alone	10	12	14	15	13	13
Psychotherapy and TCA	12	12	10	5	4	8
Psychotherapy and SSRI	24	31	39	46	48	40
TCA/SSRI	2	2	2	2	< 1	2
Psychotherapy and TCA/SSRI	5	6	9	7	4	7
Sum of guideline-compatible episodes	522	939	1353	1391	834	5039
Total episodes	1479	2211	2426	2468	1483	10,067
Percent guideline-compatible	35%	43%	56%	56%	56%	50%

psychotherapy-alone treatments, from 41 percent in 1991 to 22 percent in 1994, and then up slightly to nearly 28 percent in 1995. Second, for the medication-only treatments, SSRI alone has grown from 10 to 13 percent, even as TCA alone declined from 6 to 2 percent; the sum of the two medication-only treatments has remained relatively constant, at 15 to 16 percent. Third, most of the compositional change among treatment bundles has involved the medication-psychotherapy combination treatments. While the TCA/psychotherapy combination fell from 12 to 4 percent between 1991 and 1995, the SSRI/psychotherapy treatment share doubled, from 24 to 48 percent of all treatments. By 1995, the SSRI/psychotherapy combination had become the modal treatment bundle.

With the prices and quantities of the revised seven treatment bundles as elementary building blocks in Method 2b, we construct aggregate price indexes. Again, we found choice of formula did not greatly effect aggregate measured price change and therefore present only two price indexes: the fixed weight Laspeyres and the Fisher-Ideal index. The results are presented in the second panel of table 5.3 (Laspeyres) and table 5.4 (Fisher-Ideal).

The most striking result that immediately emerges from table 5.3 is that, over the 1991–1995 time period, the aggregate measure of price change for treatment of depression is affected substantially by the variations in inclusion/exclusion criteria between Methods 2a and 2b. With the more restrictive inclusion criteria of Method 2a, the AAGR from 1991 to 1995 is −9.1 percent, whereas in Method 2b, when the greater

number of treatment bundles and longer treatment cases are included, the 1991–1995 AAGR is but −0.6 percent. As seen in table 5.4 for the Fisher-Ideal indexes, the AAGRs are −8.5 percent (Method 2a) and −0.7 percent (Method 2b).

With the preferable Method 2b procedures, the treatment price index for acute phase major depression has hardly changed, remaining at 100 or falling slightly to around 97. These price indexes reveal an increase from 1991 to 1992, a fall in 1993, a fall again in 1994, then an increase in 1995. Differences between the fixed-weight and Fisher-Ideal indexes are relatively minor.

By comparison, over the 1991–1995 time period, the aggregate PPI for all finished goods increased about 5 percent and that for antidepressant prescription drugs increased by about 20 percent. From 1992 to 1995, the PPI for psychiatric hospital services increased by about 10 percent, while between 1994 and 1995 the overall health services PPI increased 2.4 percent. Thus, while our price indexes for the treatment of acute phase depression are either flat or very slightly falling, they all grow considerably less than the various official "discrete input" PPIs.

Critique of Method 2

These Method 2 price indexes represent a first approximation of an outcome-adjusted treatment episode price index. While the use of published treatment guidelines as a proxy measure for quality of care and expected outcomes may be preferable to ignoring outcome variations altogether, it nevertheless imposes a strong restriction on the shape of the medical treatment production function. Specifically, use of the guideline criterion implies that simply by observing input quantity combinations, one can assign each observed treatment episode either a zero (not compatible, not effective) or a one (guideline compatible, satisfactory outcome, and equal across all guideline compatible treatments)—a single-step production function. Use of the step function production relationship is very simplistic and does not make use of a great deal of clinical and medical information that is now known. Moreover, by limiting the price index to those episodes meeting guideline treatment standards, we ignore 50 percent of delivered care. We find that the share of episodes treated with guideline-level care is increasing from 35 to 55 percent from 1991 to 1995. Suggested causes of this increase include the dissemination of practice guidelines, expanded treatment choice and organizational changes—potentially real productivity gains not captured by Method 2.

VII. Method 3

Method 3 incorporates procedures that allow us to approximate more closely the effectiveness of treatments in a naturalistic setting and to recognize the variation in expected outcomes of guideline-level treatments once clinical practice deviates from the pristine world of clinical trials. We convened a panel of clinicians and researchers, provided them with detailed summaries of the clinical literature, and elicited outcomes for a wide range of treatment types and quantities. We assigned outcomes to treatment episodes based on the ratings and treatment/patient type. This allowed us to infer outcome information for episodes whose treatments do not reach the level or type recommended by guidelines and to capture variation over time in the proportion of treatments "off-frontier."[21] By integrating knowledge concerning the efficacy and effectiveness of alternative treatments based on micro clinical trial data with the judgments of clinical experts, the Method 3 price indexes better capture the effects of changing treatment practice.[22]

First we classified each episode of care previously constructed from the MarketScan data along two dimensions: type of patient and type of treatment. An example of a patient/treatment cell is women between ages 18 to 45, with no medical comorbid conditions and no recent substance abuse problems, treated with an SSRI for at least thirty days plus three or more psychotherapy visits in a specialty mental health setting.

Based on these dimensions, we identified about 200 patient-treatment cells. To reduce the burden of the expert panel's deliberations, we eliminated most of the treatment/patient cells having fewer than thirty patients over the six years. This reduction resulted in 120 treatment/patient cells covering 9,054 episodes (67 percent of ambulatory treatment episodes). Although 867 patient visits involved an ICD-9 depression diagnosis, their medical claims contained no information on any mental health treatment ($n = 816$) or there were no expenditures ($n = 51$). When these "no treatment" episodes are deleted, the number of remaining episodes dropped to 8,187. The number of rated treatment cells for the treatment bundles is given in the first column of table 5.6. We now discuss the Method 3 procedures in greater detail.

Outcomes Assessment: The Expert Panel

To obtain expected outcome probability assessments for each of the 120 treatment/patient cells, we combined research data on clinical efficacy

Table 5.6
Spending and expected outcomes of ambulatory treatments for depression

Treatment	Number of patient strata	Number (%) of episodes	Median probability of full remission	Average spending per episode	Average spending per incremental full remission
1 brief office visit	11	901 (10)	0.17	$ 53	$ 2618
2 brief office visits	1	42 (0)	0.15	95	∞
3 or more brief office visits	1	34 (0)	0.20	235	4700
1 psychotherapy visit	9	978 (11)	0.16	136	13573
2–3 psychotherapy visits	8	816 (9)	0.19	278	6930
4–9 psychotherapy visits	9	1292 (14)	0.25	558	5577
10–24 psychotherapy visits	9	637 (7)	0.34	1055	5553
TCA ≤ 30 days *and*					
1–3 psychotherapy visits	1	22 (0)	0.20	280	5585
4 or more psychotherapy visits	2	49 (1)	0.24	856	9507
TCA >30 days *and*					
No psychotherapy	2	30 (0)	0.20	280	5585
1–3 psychotherapy visits	1	25 (0)	0.30	637	4243
4 or more psychotherapy visits	2	63 (1)	0.35	864	4320
SSRI ≤ 30 days *and*					
No psychotherapy	2	65 (1)	0.20	173	3447
1–3 psychotherapy visits	2	87 (1)	0.20	304	6070
4 or more psychotherapy visits	3	147 (2)	0.25	877	8765
4 or more psychotherapy visits, some anxiolytics	2	18 (0)	0.32	837	4924
Trazodone ≤ 30 days, some anxiolytics, some psychotherapy	1	19 (0)	0.20	756	15116

Table 5.6 (continued)

Treatment	Number of patient strata	Number (%) of episodes	Median probability of full remission	Average spending per episode	Average spending per incremental full remission
SSRI >30 days *and*					
No psychotherapy	11	552 (6)	0.28	$ 308	$ 2367
1–3 psychotherapy visits	7	473 (5)	0.29	504	3593
4 or more psychotherapy visits	9	801 (9)	0.35	1050	5249
No psychotherapy, some anxiolytics	1	20 (0)	0.32	315	1853
1–3 psychotherapy visits, some anxiolytics	1	36 (0)	0.35	553	2765
Trazodone ≤ 30 days, some anxiolytics, some psychotherapy	1	35 (0)	0.30	875	5827
Trazodone > 30 days, some anxiolytics, some psychotherapy	2	61 (1)	0.35	1167	5833
Heterocyclics > 30 days, some anxiolytics, some psychotherapy	1	39 (0)	0.28	557	4284
Lithium alone	1	20 (0)	0.20	538	10753
Lithium plus antidepressants	2	67 (0)	0.26	820	7454
Antipsychotics alone	1	15 (0)	0.15	476	∞
Unspecified mental health treatment	8	894 (10)	0.16	757	75690
No mental health treatment	9	816 (9)	0.15[a]	—	—
Total	120	9054	0.23	473	5909

[a]Median probability of being depression-free was 0.18 but was set to the minimum value of 0.15 for calculating incremental spending.

and effectiveness with expert clinical opinion.[23] We asked ten clinicians to rate outcomes from these treatments in terms of changes in the Hamilton Depression Rating Scale (HDRS). The elicitation process consisted of three distinct stages. In the first stage, a psychiatrist provided a summary of published results of treatment efficacy and effectiveness for each of the treatment/patient cells. In the second stage, an elicitation of outcome distributions was conducted by mail for each of the 120 treatment/patient cells.[24] Finally, a face-to-face panel meeting was convened in January 1999, and each expert was asked to rate again those cells having a substantial disagreement among experts. For our purposes, the relevant outcome was the share of patients depression-free as measured by the HDRS after sixteen weeks of treatment. Panel consensus emerged at the end of the panel meeting.

The weighted average of the median outcome results from the expert panel elicitation process for each treatment (averaged over patient type) and average spending from the MarketScan data are reported in columns 3 and 4, respectively, of table 5.6.[25] The ratings indicate substantial variation in several of the treatments previously assumed in Method 2 to have equivalent efficacy. The average expected outcome rating of these guideline-level treatments ranged from 20 (TCA alone) to 35 percent (psychotherapy plus SSRIs). This range suggests that the quality effect of changes in treatments over time, as indicated in table 5.5, may affect the price index. Moreover, similar to results in clinical trials, the ratings indicate that clinicians believe 15 out of 100 patients will spontaneously remit, even in the absence of recognized treatments for depression.

In implementing an outcome-adjusted price index, the relevant outcome increment is the likelihood of symptom reduction relative to the "waiting list" or "no treatment" outcome. Therefore, we incorporate the fact that the no-treatment option ("watchful waiting" or "waiting list") frequently results in spontaneous remission of depressive symptoms and focus on the *incremental* gains in outcome obtained from treatment over and above waiting list. Reported in the final column of table 5.6 is the average *incremental* spending per full remission.

To gauge the effect of incorporating more refined outcome adjustments, we construct two distinct series of price indexes. First, we construct price indexes similar to those in Method 2b but increase the number of episodes used to construct the price index by including all patient/treatment cells rated by the panel. This change increases the

share of episodes included in price index calculations by 40 percent (from 50 to 70 percent of all episodes). Because measures of outcome variation are not incorporated, all cells rated by the panel are treated as having equivalent outcome.

In the preferred Method 3 price index, we combine the ratings of the expert panel's outcome assessments with changes over time in the shares of observed treatment modalities (and their costs). Thus, this price index measures changes in the price per incremental remission. Once again, we limit our discussion to the Laspeyres and Fisher-Ideal indexes.[26]

Results

The two series of price indexes described above are noted in the third panel of tables 5.3 and 5.4. Limiting the price index to the rated episodes (Method 3, row 1), without reference to outcome, we observe that the treatment price index falls from 100 in 1991 to 95.0 (Laspeyres), and 97.2 (Fisher-Ideal) in 1996. These declines look remarkably similar to those recorded using Method 2b. Once the rated outcomes from the expert panel are incorporated, however, a somewhat different pattern emerges. For both the Laspeyres (table 5.3) and Fisher-Ideal (table 5.4) indexes, prices show a decline to 92 in 1992, an increase back up to about 100 in 1993, a small increase to 101 to 102 in 1994, a larger increase to 108 in 1995, and then a decrease to about 103 to 104 in 1996. Because we observed an increase in the quality of care over time, the sharp increases in outcomes-adjusted prices in 1993 and 1995 are surprising.

In considering this surprising increase, we note that during the 1991 through 1996 period, increased levels of management have been exercised over the mental health benefit. One implication of this trend is that the patient population being treated may have been changing (along with the mix of treatment bundles), thereby affecting both expected outcome and cost. Because our expected outcomes are assigned based on both treatment and patient type, changes in the mix of patients will affect the price per remission. For example, the panel rated patients with comorbid substance abuse to have lower expected outcomes than patients without comorbid substance abuse receiving the same treatment. These changes in the patient mix over time in our sample are not incorporated in the Method 3 price index.

VIII. Method 4

To account for the effect of changing patient mix on computed price indexes, we employ hedonic-like equations. In Method 4, we use regression analysis to account for varying patient characteristics when calculating the price index.

We have specified and estimated several such models, based on data from the 8,187 rated treatment episodes.[27] We delineate eight patient categories, depending on whether medical comorbidity is present; whether male; if female, whether over age 50; and whether there is comorbid substance abuse. We estimated hedonic equations for the price per expected full remission. The left-hand variable is the natural log of spending per episode of treatment. The right-hand variables include the probability of a full remission associated with the patient's treatment and patient type, dummy variables for seven of eight patient categories, and year dummies (1991 is excluded).[28]

As expected, variations in patient categories were found to have significant and substantial effects on treatment costs. The coefficient on remission probability is positive and highly significant. The resulting price index is presented in the last row of table 5.3. The magnitude of the reductions from 1991 to 1996 is 12.8 percent in price per remission, with an AAGR of −2.7 percent. The 1991 to 1996 differences are significantly different from zero at the ($p < 0.05$) level, although in some cases the intervening year estimates are not very precise ($p < 0.10$). The differences between the hedonic and traditional price indexes appear to result, therefore, from a changing and increasingly complex mix of patients, along with changes in treatment bundles over the six-year period.

It is important to note that the hedonic regressions suggest a rather different set of price movements than the traditional price indexes. The main reason for the difference is that variables describing the composition of cases are included in the hedonic regression but are not properly controlled for in the traditional price index calculations. The mix of patients has been changing over time. To illustrate this point, consider depressed patients with medical comorbidities. In 1991, about 13 percent of episodes involved comorbid medical diagnoses, but by 1995 nearly 23 percent of episodes involved such conditions, and in 1996 this share fell to 19 percent. Comorbidities are known to complicate treatment and make it more costly. Thus, when the hedonic model controls for the

changing patient composition of the treated population, as well as changes in expected outcomes, the price indexes show declines in the nominal price of an expected remission.

IX. Decomposing Price Index Discrepancies

The variations in procedures used in Method 1 through Method 4 to construct price indexes for the treatment of depression are summarized in table 5.1. The price indexes presented in tables 5.3 and 5.4 attempt to incorporate several conceptual improvements suggested by economists concerning the measurement of medical care price changes. Differences between these indexes and official government-produced price indexes emerge from several sources. In this section, we evaluate the relative importance of the sources of these discrepancies. We begin with the two BLS-like price indexes constructed for discrete inputs for the care of depression, presented in the top panels of tables 5.3 and 5.4. These price indexes mimic methods used in the construction of official price indexes but use as inputs prices and quantities from the Medstat MarketScan data. We incrementally add improvements, assessing the resulting change in the price index at each of three methodological steps.

To construct a BLS-like health care PPI for the treatment of depression, we take a weighted average of the estimate of the price index for antidepressant drugs and the estimate of the price index for physicians' services. The relevant weights are derived from the 1991 MarketScan data as the share of total expenditures attributed to each of these categories. The resulting weights are 9 percent and 91 percent for antidepressants and physicians' services, respectively. We first combine the two indexes calculated using the fixed quantity weight Laspeyres formula using these relative weights. As indicated in line 1 of table 5.7, combining these indexes indicates a positive AAGR of 2.6 percent in the price of health care used for the treatment of depression.

We repeat this calculation using the same weighting scheme but combine the two price indexes calculated using the Fisher-Ideal index formula. These results are noted in line 2 of table 5.7. Comparing the fixed weight index with an index that includes more recent data in the calculation, we find the AAGR declines from 2.6 percent to 2.0 percent. Compared to the Laspeyres formula, the Fisher-Ideal formula indicates the same price increase for antidepressant drugs but a decline in the AAGR by 28 percent for physicians' services. Therefore, the smaller

Table 5.7
Summary of price index calculations, 1991–1996

	AAGR
(1) BLS-like medical care PPI for depression using MarketScan data—Laspeyres formula	+2.6
(2) BLS-like medical care PPI for depression using MarketScan data—Fisher-Ideal formula	+2.0
(3) Treatment-episode approach	−0.7
(4) Incorporating expected outcome of treatments, adjusting for patient characteristics	−2.7
Total bias due to methodology	−5.3

price increase found when using the Fisher-Ideal indexes to compute the BLS-like medical care PPI is due exclusively to more current weights in the physicians' services index. The similarity of the antidepressant drug price index under the two formulas might suggest that the changing mix of treatments over time within a class (e.g., from one drug to another) is not important for the measurement of price changes. Although not reported here, a closer look at the data underlying these calculations indicates some trends. We note a change in the mix of drugs provided over time and differences in price change among the drugs. Some have very small increases or a constant price, while others have quite large increases. Therefore, the absence of a difference in price change when using more current quantity weights is not due to systematic similarities in pricing trends among the drugs. While the index number formula had a rather small effect on the BLS-like medical care PPI for depression, this does not suggest this is likely to be the case for changes in the price of treatment for other diseases.

In the next step of our decomposition, we compare a discrete input approach to an episode of treatment approach. This comparison yields a much different view of price movements over the same time period. Line 3 of table 5.7 presents the AAGR calculated using a treatment-episode approach. Note that variations in the expected outcome of treatments are not yet incorporated in this calculation. The negative AAGR of less than 1 percent indicates a slight *decline* in price over this time period. The difference between line 2 and line 3 may reflect changes across treatment classes (e.g., a move from outpatient services to drug treatments). If goods or services that are substitutes are not in the same BLS item strata, calculating price movement for item strata separately will not capture substitution between the item strata. At least in the

case of treatment for depression, this fact may play an important role in price change, for there has been significant substitution away from psychotherapy to drug treatments. A second potential explanation of the difference in price movements found using the treatment episode as the unit of analysis is a decline in the quantity of treatment inputs within treatment episodes. In later years, patients may be receiving fewer psychotherapy visits or days of drug treatment within an episode.

In the final step indicated in line 4 of table 5.7, we present a price index that incorporates both changes in expected outcome and in patient characteristics based on expected outcome measures from an expert panel. This modification to the price index methodology further reduces the estimate of price changes, indicating a *negative* AAGR of 2.7 percent over the 1991–1996 time period. This figure suggests that improvements in expected outcomes resulting from changes in treatments prescribed over this time period had a significant effect on productivity in delivering health care services to treat depression.

This decomposition analysis indicates that the discrepancies found between our preferred methodology and the BLS methodology can be attributed mainly to moving to a treatment-episode approach and to failure to account for changes in expected outcome. This suggests, at least in the case of treatment for depression, that recent modifications to BLS methodology such as more frequent updating of weights and the use of a geometric mean formula, which do not address these discrepancies, will have little effect on more reliably measuring price changes over time in treating depression.

Our analysis does not enable us to predict the result of the BLS' changes for the reliability of measuring changes in the cost of treatment for other diseases; these BLS changes may in fact be important in reliably measuring price changes for the treatment of other diseases. Our results also suggest that more research may be needed on the impact of the classification of goods and services to item strata and on the type of classification system that will best capture the full impact of substitution among treatment classes over time.

X. Conclusion

We found the construction of various price indexes for the treatment of depression to be instructive and useful. Several general lessons emerge

from this research. First, we found that moving to an episode-of-care approach matters for the measurement of price changes. Conceptually this step is important because it takes us closer to what patients are actually buying when they consume medical care—episodes of treatment versus individual health care inputs. Second, we learned that starting to incorporate at least some measure of outcomes is important. Including outcomes also takes us closer to what individuals are actually buying when they purchase health care. Finally, if we contrast our preferred price index to traditional medical care price indexes, we find that prices decline and productivity increases, rather than the opposite. However, constructing episode-based, outcome-adjusted price indexes is an extraordinarily difficult and cumbersome task. For public policy and evaluation purposes, construction of such indexes is very important. However, it is not necessarily practical or sensible for the Bureau of Labor Statistics to calculate such an index on a monthly basis. When public policy analysts and researchers evaluate and interpret changes over time in the National Health Accounts, it is important that they bear in mind that expenditures are the product of price times quantity. If price indexes fail to portray reliably the changing use of medical inputs to treat episodes of illness, and if outcomes variation and patient complexity are not taken into account, it becomes difficult if not impossible to identify the sources of expenditure changes. How frequently, how detailed, and by whom such medical care price indexes should be constructed is an issue meriting urgent attention.

Notes

Funding from the John D. and Catherine T. MacArthur Foundation, the NIMH, and the National Science Foundation is gratefully acknowledged.

1. Kessler, McGonagle, Zhao et al. (1994), p. 12.

2. Kendler, Walters, Neale et al. (1995).

3. Kessler, McGonagle, Swartz et al. (1993), p. 91.

4. See Busch, Frank, and Berndt (1996).

5. The Depression Guidelines Panel (1993) provides a summary and interpretation of the evidence on this point.

6. See American Psychiatric Association (1993).

7. Keller et al. (2000).

8. Levit, Lazenby, Braden et al. (1998), and Ford and Ginsburg (1997).

9. For further discussion, see Berndt, Cockburn, and Griliches (1996), Griliches and Cockburn (1994), Kanoza (1996), and Kelly (1997).

10. All procedure codes comprising at least ½ of 1 percent of all such outpatient claims were included.

11. See, for example, Diewert (1976, 1981).

12. For a more detailed explanation of these methods see Frank et al. 1998a and Frank et al. 1998b.

13. We thank Marcella-Horwitz Lennon, Howard Goldman, and Alyssa Busch for their clinical expertise.

14. Medicode (1998).

15. For discussion of defining episodes of care, see Keeler et al. (1986) and Wingert et al. (1995).

16. We count days without treatment only after the number of days of supply in a drug prescription has been exhausted, thereby assuming full compliance with the daily recommended dosage.

17. A significant portion of inpatient episodes have unspecified outpatient follow-up, thereby limiting that avenue for identifying treatments.

18. For a discussion, see Mechanic (1989) and McGuire (1989).

19. These were brand name drugs Effexor and Serzone.

20. See Crown, Obenchain, Englehart et al. (1996) and Wilde and Benfield (1998) for evidence on differential TCA-SSRI compliance among patients.

21. There is a large amount of literature on the cross-sectional variation in medical treatments for other conditions and illnesses. See, for example, Phelps (1992) and Phelps and Mooney (1993). On the effects of cost containment and risk-sharing on choice of treatment, see Berndt, Frank, and McGuire (1997); Goldman, McCulloch, and Sturm (1998); and Ma and McGuire (1998).

22. See also Frank et al. 1999.

23. See Normand, Frank and McGuire (1999).

24. Experts were directed: "Using your best clinical judgment, consider an average group of 100 patients presenting with major depressive disorder with an entry HDRS score of 22 in 1998 to an average specialty mental health or primary care physician. After sixteen weeks of the indicated treatment, how many patients will most likely fall into each of the following categories: depression-free, mildly depressed, moderately depressed, or no change? When completing the survey, please keep in mind the impact of compliance and the average circumstances of routine practice on symptom reductions."

25. This table is reproduced from Normand, Frank and McGuire (1999) and is based on a slightly smaller sample of episodes.

26. For the results of additional price index calculations, see Berndt et al. (2000a).

27. For a discussion of index number and hedonic methods to control for quality in the medical care context and elsewhere, see Cockburn and Anis (2001), Gilbert (1961, 1962),

Griliches (1962), and Reder (1969). Price indexes with heterogeneous purchasers are considered by Diewert (1981), Fisher and Griliches (1995), and Griliches and Cockburn (1994).

28. Full regression results appear in Berndt et al. (2000a).

References

American Psychiatric Association (1993). "Practice Guidelines for Major Depressive Disorder in Adults," *American Journal of Psychiatry*, 150:4, Supplement, April, 1–26.

American Psychiatric Association (1994). *Diagnostic and Statistic Manual of Mental Disorders*, 4th Edition, Washington D.C.

Berndt, Ernst R., Anupa Bir, Susan H. Busch, Richard G. Frank, and Sharon-Lise T. Normand (2000a). "The Medical Treatment of Depression, 1991–1996: Productive Inefficiency: Expected Outcome Variations and Price Indexes." NBER Working Paper.

Berndt, Ernst R., Iain Cockburn, and Zvi Griliches (1996). "Pharmaceutical Innovations and Market Dynamics: Tracking Effects on Price Indexes for Antidepressant Drugs," Brookings Papers on Economic Activity: Microeconomics 1996, 133–188.

Berndt, Ernst R., David Cutler, Richard G. Frank, Zvi Griliches, Joseph Newhouse, and Jack Triplett (2000b). "Medical Care Prices and Output," Chapter 3 in A. J. Culyer and J. P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1A, Amsterdam: Elsevier Science, pp. 119–180.

Berndt, Ernst R., Richard G. Frank, and Thomas G. McGuire (1997). "Alternative Insurance Arrangements and the Treatment of Depression: What Are the Facts?" *American Journal of Managed Care*, 3:2, February, 243–250.

Busch, Susan H., Richard G. Frank, and Ernst R. Berndt (1996). "Effectiveness, Efficacy and Price Indexes for Depression: A Review of the Literature," unpublished manuscript, Boston: Harvard Medical School, Department of Health Care Policy, December.

Cockburn, Iain M., and Aslam Anis (2001). "Hedonic Analysis of Arthritis Drugs," paper given at NBER-CRIW Conference on Medical Output and Productivity, June 12–13, 1998. Forthcoming in David Cutler and Ernst R. Berndt, eds., *Medical Care Output and Productivity*, Chicago: University of Chicago Press.

Crown, William E., Robert L. Obenchain, Luella Englehart, Tamra Lair, Don P. Buesching, and Thomas W. Croghan (1996). "Application of Sample Selection Models to Outcomes Research: The Case of Evaluating Effects of Antidepressant Therapy on Resource Utilization," unpublished manuscript, Cambridge, MA: The MEDSTAT Group, January.

Depression Guidelines Panel (1993). *Depression in Primary Care*, Vol. 2, Rockville, MD: U.S. Department of Health and Human Services, AHCPR publication 93-0551.

Diewert, W. Erwin (1976). "Exact and Superlative Index Numbers," *Journal of Econometrics*, 4:2, 115–145.

Diewert, W. Erwin (1981). "The Economic Theory of Index Numbers: A Survey," in Angus Deaton, ed., *Essays in the Theory and Measurement of Consumer Behavior in Honour of Sir Richard Stone*, Cambridge: Cambridge University Press, 163–208.

Fisher, Franklin M., and Zvi Griliches (1995). "Aggregate Price Indexes, New Goods, and Generics," *Quarterly Journal of Economics*, Vol. 110, No. 1, February, pp. 229–244.

Ford, Ina Kay, and Daniel H. Ginsburg (1997). "Medical Care in the CPI," paper given in presentation to the NBER Summer Institute, Franco-American Seminar, Cambridge, MA, July 23.

Frank, Richard G., Ernst R. Berndt, and Susan H. Busch (1998a). "Price Indexes for the Treatment of Depression," in Jack E. Triplett, ed., *Measuring the Prices of Medical Treatments*, Washington D.C.: The Brookings Institution.

Frank, Richard G., Susan H. Busch, and Ernst R. Berndt (1998b). "Measuring Prices and Quantities of Treatment for Depression," *American Economic Review*, 88:2, May, 106–111.

Frank, Richard G., Thomas G. McGuire, Sharon-Lise Normand, and Howard H. Goldman (1999). "The Value of Mental Health Care at the System Level: The Case of Treating Depression," *Health Affairs*, Vol. 18, No. 5, September/October, pp. 71–88.

Gilbert, Milton (1961). "The Problem of Quality Change and Index Numbers," *Monthly Labor Review*, 84:9, September, 992–997.

Gilbert, Milton (1962). "Quality Change and Index Numbers: A Reply," *Monthly Labor Review*, 85:5, May, 544–545.

Goldman, William, Joyce McCulloch, and Roland Sturm (1998). "Costs and Use of Mental Health Services Before and After Managed Care," *Health Affairs*, 17:2, March/April, 40–52.

Griliches, Zvi (1962). "Quality Change and Index Numbers: A Critique," *Monthly Labor Review*, 85:5, 532–544.

Griliches, Zvi, and Iain Cockburn (1994). "Generics and New Goods in Pharmaceutical Price Indexes," *American Economic Review*, 84:5, December, 1213–1232.

Kanoza, Douglas (1996). "Supplemental Sampling in the PPI Pharmaceuticals Index," *Producer Price Indexes Detailed Price Report*, January, 8–10.

Keeler, Emmett B., Kenneth B. Wells, Willard G. Manning, J. David Rumpel, and Janet M. Hanley (1986). "The Demand for Episodes of Mental Health Services," Santa Monica, CA: RAND Report R-3432-NIMH, October.

Keller, Martin B., James P. McCullough, Daniel N. Klein, Bruce Arnow, David L. Dunner, Alan J. Gelenberg, John C. Markowitz, Charles B. Nemeroff, James M. Russell, Michael E. Thase, Madhukar H. Trivedi, John Zajecka, Janice A. Blalock, Francis E. Borian, Charles DeBattista, Jan Fawcett, Robert M. A. Hirschfeld, Darlene N. Jody, Gabor Keitner, James H. Kocsis, Lorrin M. Koran, Susan G. Kornstein, Rachel Manber, Ivan Miller, Philip T. Ninan, Barbara Rothbaum, A. John Rush, Alan F. Schatzberg, and Dina Vivian (2000). "A Comparison of Nefazodone, the Cognitive Behavioral Analysis System of Psychotherapy and Their Combination for Treatment of Chronic Depression," *New England Journal of Medicine*, Vol. 342, No. 20, May 18, 2000, pp. 1462–1470.

Kelly, Gregory G. (1997). "Improving the PPI Samples for Prescription Pharmaceuticals," *Monthly Labor Review*, 120:10, October, 10–17.

Kendler, Kenneth S., Ellen E. Walters, Michael C. Neale et al. (1995). "The Structure of the Genetic and Environmental Risk Factors for Six Major Psychiatric Disorders in Women. Phobia, Generalized Anxiety Disorder, Panic Disorder, Bulimia, Major Depression, and Alcoholism," *Archives of General Psychiatry*, 52:5, May 1995, pp. 374–383.

Kessler, Ronald C., Katherine A. McGonagle, Marvin Swartz, Dan G. Blazer, and Christopher B. Nelson (1993). "Sex and Depression in the National Comorbidity Survey I: Lifetime Prevalence, Chronicity and Recurrence," *Journal of Affective Disorders*, 29:85–96.

Kessler, Ronald C., Katherine A. McGonagle, Shanyang Zhao, Christopher B. Nelson, Michael Hughes, Suzann Eshlerman, Hans-Ulrich Wittchen, and Kenneth S. Kendler (1994). "Lifetime and Twelve-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results from the National Comorbidity Survey," *Archives of General Psychiatry*, 51:1, 8–19.

Levit, Katharine R., Helen C. Lazenby, Brandley R. Braden, and the National Health Accounts Team (1998). "National Health Spending Trends in 1996," *Health Affairs*, 17:1, January/February, 35–51.

Ma, Ching-to Albert, and Thomas G. McGuire (1998). "Costs and Incentives in a Behavioral Health Carve-Out," *Health Affairs*, 17:2, March/April, 53–69.

McGuire, Thomas G. (1989). "Financing and Reimbursement for Mental Health Services," in Carl A. Taube, David Mechanic, and Ann A. Hohmann, eds., *The Future of Mental Health Services Research*, Rockville, MD: U.S. Department of Health and Human Services, 87–112.

Mechanic, David (1989). "The Evolution of Mental Health Services and Mental Health Services Research," in Carl A. Taube, David Mechanic, and Ann A. Hohmann, eds., *The Future of Mental Health Services Research*, Rockville, MD: U.S. Department of Health and Human Services, 1–8.

Medicode (1998). *International Classification of Diseases, 9th Revision, Clinical Modification*, 5th Edition, Salt Lake City: Medicode.

Normand, Sharon-Lise T., Richard G. Frank, and Thomas G. McGuire (1999). "Elicitation in Observational Studies: Estimating the Value of Acute Treatments for Depression," Boston, MA: Harvard Medical School, draft manuscript, December 26.

Phelps, Charles E. (1992). "Diffusion of Information in Medical Care," *Journal of Economic Perspectives*, Vol. 6, No. 3, Summer, 23–42.

Phelps, Charles E., and Cathleen Mooney (1993). "Variations in Medical Practice Use: Causes and Consequences," in Richard J. Arnould, Robert F. Rich, and Wiliam D. White, eds., *Competitive Approaches to Health Care Reform*, Washington D.C.: Urban Institute Press.

Reder, Melvin W. (1969). "Some Problems in the Measurement of Productivity in the Medical Care Industry," in Victor R. Fuchs, ed., *Production and Productivity in the Service Industries*, NBER Studies in Income and Wealth, Vol. 34, New York: Columbia University Press, pp. 95–131.

Scitovsky, Anne A. (1967). "Changes in the Costs of Treatment of Selected Illnesses, 1951–65," *American Economic Review*, 57:5, December, 1182–1195.

Wilde, Michelle I., and Paul Benfield (1998). "Fluoxetine: A Pharmacoeconomic Review of Its Use in Depression," *Pharmacoeconomics*, 13:5 (1), May, 543–561.

Wingert, Terence D., John E. Kralewski, Tammie J. Lindquist, and David J. Knutson (1995). "Constructing Episodes of Care from Encounter and Claims Data: Some Methodological Issues," *Inquiry*, Vol. 32, Winter, pp. 430–443.